

# Balancing Covariates via Propensity Score Weighting

Fan Li      Kari Lock Morgan      Alan M. Zaslavsky<sup>1</sup>

## ABSTRACT

Covariate balance is crucial for an unconfounded descriptive or causal comparison. However, lack of balance is common in observational studies. This article considers weighting strategies for balancing covariates. We define a general class of weights—the balancing weights—that balance the weighted distributions of the covariates between treatment groups. These weights incorporate the propensity score to weight each group to an analyst-selected target population. This class unifies existing weighting methods, including commonly used weights such as the inverse-probability weights as special cases. General large-sample results on nonparametric estimation based on these weights are derived. We further propose a new weighting scheme, the overlap weights, in which each unit’s weight is proportional to the probability of that unit being assigned to the opposite group. The overlap weights are bounded, minimize the asymptotic variance of the weighted average treatment effect among the class of balancing weights, and possess a desirable small-sample exact balance property. Two applications illustrate this method and compare it with other approaches.

**KEY WORDS:** balance, causal inference, confounding, controlled comparison, overlap, propensity score, weighting.

---

<sup>1</sup>© 2014 Fan Li, Kari Lock Morgan, Alan M. Zaslavsky. All Rights Reserved. Fan Li is assistant professor, Department of Statistical Science, Duke University, Durham, NC, 27705 (email: fli@stat.duke.edu); Kari Lock Morgan is assistant professor, Department of Statistics, Penn State University, University Park, PA 16802 (email: klm47@psu.edu); Alan M. Zaslavsky is professor, Department of Health Care Policy, Harvard Medical School, Boston, MA 02115 (email: zaslavsk@hcp.med.harvard.edu). The authors thank Dylan Small for sharing the programming code for the RHC study. Li and Morgan’s research is partially funded by NSF-SES grant 1424688.

# 1 Introduction

Unconfounded comparison between groups is a central goal in many observational studies. Comparative effectiveness studies aim to estimate the causal effect of a treatment or intervention unconfounded by differences between characteristics of those assigned to the treatment and control conditions under current practice. Noncausal descriptive studies often concern a controlled and unconfounded comparison of two populations, such as comparing outcomes among populations of different races or of cohorts in different years while giving the comparison groups similar distributions of some important covariates. Whether the purpose of the study is causal or descriptive, comparisons between groups can be biased when the groups lack balance, that is, have substantially different distributions of relevant covariates.

Standard parametric adjustment by regression is often sensitive to model misspecification (Rubin, 1979) when groups differ greatly in observed characteristics. Two widely used non-parametric balancing strategies are matching and weighting. Matching links “similar” cases with respect to confounders in observed samples from two groups according to some distance measure (e.g., Mahalanobis distance or differences in propensity score) and the comparison is made based on the matched sample. Weighting methods apply weights to the entire sample of each treatment group to match the covariate distribution of a target population, and the comparison is made between the weighted outcomes. Matching is designed to create local balance for a subset of the observed sample, whereas weighting is designed to create global balance for the target population. In this paper we focus on weighting, and will later show that many-to-many matching can be regarded as a form of weighting.

The current literature on weighting (e.g. Robins and Rotnitzky, 1995; Hahn, 1998; Hirano and Imbens, 2001; Hirano *et al.*, 2003; Imbens, 2004; Crump *et al.*, 2009) largely focuses on the Horvitz-Thompson (HT) weights, also known as inverse-probability weights, originating from survey research (Horvitz and Thompson, 1952). The HT weight for each unit is the inverse of

the probability of that unit being assigned to the observed group. The HT weights correspond to estimands defined on the population represented by the combined treatment and control groups, such as the average treatment effect (ATE) in causal studies. Others have proposed alternative weighting schemes also targeted at some well-defined subpopulations, such as the average treatment effect for the treated (ATT) or for the control group (e.g. Hirano and Imbens, 2001). An operational drawback of inverse-probability weights is that extreme probabilities (close to 0 or 1) introduce extremely large weights, which can dominate the estimates and result in poor balance and very large variance. A common practice is to truncate the extreme weights based on an arbitrary cutoff point, either discarding the sample units with weights beyond the cutoff or setting these weights to the cutoff value. Such *ad hoc* procedures lack theoretical basis for the choice of target population.

In this article, we first define the class of *balancing weights* — those that balance the distributions of covariates between comparison groups — and show they can be defined for any pre-specified target population. Special cases include the several aforementioned weights. Within this class of weights, we then introduce the *overlap weights*, which weight each unit proportional to its probability of assignment to the opposite group. Unlike the inverse probability weights, the overlap weights have the advantage of being bounded between 0 and 1. Under mild conditions, the overlap weights minimize the asymptotic variance of the nonparametric estimate of the weighted average treatment effect within the class of balancing weights. We further show that the overlap weights estimated from a logistic model lead to exact (small-sample) balance between the groups in the means of the covariates. The target population corresponding to the overlap weights has a clinical or policy interpretation as the part of the population which currently receives either treatment in substantial proportions. Our work adds to a recent strand of literature (Rosenbaum, 2012; Crump *et al.*, 2009; Traskin and Small, 2011) that focuses on comparisons in a subpopulation with sufficient overlap. However, our approach differs in that

all units are weighted continuously, rather than making binary decisions to include or exclude units from the subsample being analyzed.

Section 2 introduces the general framework of balancing weights and defines corresponding estimands. Section 3 presents two large-sample properties of the nonparametric weighting estimators, following which we propose the overlap weight in Section 4 and discuss its theoretical and practical properties. In Section 5, we illustrate the proposed method through two applications. Section 6 concludes.

## 2 Balancing Weights

Consider a sample or finite population of  $N$  units, each belonging to one of two groups for which covariate-balanced comparisons are of interest, possibly defined by a treatment. Let  $Z_i = z$  be the binary variable indicating membership in groups that may be labeled treatment ( $z = 1$ ) and control ( $z = 0$ ). For each unit, an outcome  $Y_i$  and a set of covariates  $X_i = (X_{i1}, \dots, X_{iK})$  are observed. The propensity score (Rosenbaum and Rubin, 1983) is the probability of assignment to the treatment group given the covariates,  $e(x) = \Pr(Z_i = 1 | X_i = x)$ .

In descriptive comparisons, “assignment” is to a nonmanipulable state defining membership in one of two groups, and a common objective is to evaluate the average difference in the outcome in two groups with balanced distributions of covariates. We first define the conditional average controlled difference (ACD) (Li *et al.*, 2013) for a given  $x$  as,

$$\tau(x) \equiv \mathbb{E}(Y | Z = 1, X = x) - \mathbb{E}(Y | Z = 0, X = x). \quad (1)$$

For causal comparisons, we adopt the potential outcome framework (Rubin, 1974, 1978). Assuming the standard Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980), which states that the potential outcomes for each unit are unaffected by the treatment assignments of other units, each unit has two potential outcomes  $Y_i(z)$  for  $z = 0, 1$ , corresponding

to each possible treatment level, and only one of the two is observed:  $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$ . Under the *unconfoundedness* assumption, that is,  $\{Y(0), Y(1)\} \perp Z | X$ , we have  $\Pr(Y(z)|X) = \Pr(Y|X, Z = z)$  for  $z = 0, 1$ , so  $\tau(x)$  in (1) is a causal estimand—the average treatment effect (ATE) conditional on  $x$ :

$$\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]. \quad (2)$$

Both comparisons require the *probabilistic assignment assumption*,  $0 < e(X) < 1$ , which states that the study population is restricted to values of covariates for which there can be both control and treated units. Assignment mechanisms assuming both probabilistic assignment and unconfoundedness are called *strongly ignorable* (Rosenbaum and Rubin, 1983).

Typically in either descriptive or causal comparisons the (potential) outcomes are compared not for a single  $x$  but rather averaged over a hypothesized target distribution of the covariates. Assume the marginal density of the covariates  $X$ ,  $f(x)$ , exists, with respect to a base measure  $\mu$  (a product of counting measure with respect to categorical variables and Lebesgue measure for continuous variables). We then represent the target population density by  $f(x)h(x)$ , where  $h(\cdot)$  is pre-specified function of  $x$ , and define a general class of estimands by the expectation of the conditional ACD or ATE  $\tau(x)$  over the target population:

$$\tau_h \equiv \frac{\int \tau(dx) f(x) h(x) \mu(dx)}{\int f(x) h(x) \mu(dx)}. \quad (3)$$

For descriptive comparisons,  $\tau_h$  is the weighted ACD, and for causal comparisons it is the weighted average treatment effect (WATE) (Hirano *et al.*, 2003), the term we use henceforth for either case.

Let  $f_z(x) = \Pr(X = x | Z = z)$  be the density of  $X$  in the  $Z = z$  group, then

$$f_1(x) \propto f(x)e(x), \quad \text{and} \quad f_0(x) \propto f(x)(1 - e(x)).$$

For a given  $h(x)$ , to estimate  $\tau_h$ , we can weight  $f_z(x)$  to the target population using the following weights (proportional up to a normalizing constant):

$$\begin{cases} w_1(x) \propto \frac{f(x)h(x)}{f(x)e(x)} = \frac{h(x)}{e(x)}, & \text{for } Z = 1, \\ w_0(x) \propto \frac{f(x)h(x)}{f(x)(1-e(x))} = \frac{h(x)}{1-e(x)}, & \text{for } Z = 0. \end{cases} \quad (4)$$

We call this class of weights defined in (4) the *balancing weights* because they balance the weighted distributions of the covariates between comparison groups:

$$f_1(x)w_1(x) = f_0(x)w_0(x) = f(x)h(x). \quad (5)$$

The function  $h$  can take any form, and all weights that balance the covariate distributions between groups can be specified within this class.

Specification of  $h$  defines the target population and estimands and determines the weights. Statistical, scientific and policy considerations all may come into play. When  $h(x) = 1$ , the corresponding target population  $f(x)$  is the combined (treated and control) population, the weights  $(w_1, w_0)$  are the HT weights  $(1/e(x), 1/(1 - e(x)))$ , and the estimand is the ATE for the combined population,  $\tau^{\text{ATE}} = \mathbb{E}[Y(1) - Y(0)]$ . When  $h(x) = e(x)$ , the target population is the treated subpopulation, the weights are  $(1, e(x)/(1 - e(x)))$ , and the estimand is the average treatment effect for the treated (ATT),  $\tau^{\text{ATT}} = \mathbb{E}[Y(1) - Y(0)|Z = 1]$ . When  $h(x) = 1 - e(x)$ , the target population is the control subpopulation, the weights are  $((1 - e(x))/e(x), 1)$ , and the estimand is the average treatment effect for the control (ATC),  $\tau^{\text{ATC}} = \mathbb{E}[Y(1) - Y(0)|Z = 0]$ . By choosing  $h$  from the class of indicator functions, one can define the ATE for truncated subpopulations of substantive interest or with desirable theoretical properties. For example, Crump *et al.* (2009) proposed to use  $h(x) = \mathbf{1}(\alpha < e(x) < 1 - \alpha)$  with a pre-specified  $\alpha \in (0, 1/2)$  that defines a subpopulation with sufficient overlap of covariates between two groups. Formulation (4) also allows choices of  $h$  that are not functions of the propensity score. For instance, if one covariate  $x$  is age, then setting  $h = \mathbf{1}(a < x_{\text{age}} < b)$  defines a target population of those with ages between  $a$  and  $b$ . These examples are summarized in Table 1.

Table 1: Examples of balancing weights and corresponding target population and estimand under different  $h$ .

target population	$h(x)$	estimand	weight $(w_1, w_0)$
combined	1	ATE	$\left(\frac{1}{e(x)}, \frac{1}{1-e(x)}\right)$ [HT]
treated	$e(x)$	ATT	$\left(1, \frac{e(x)}{1-e(x)}\right)$
control	$1 - e(x)$	ATC	$\left(\frac{1-e(x)}{e(x)}, 1\right)$
truncated combined	$\mathbf{1}(\alpha < e(x) < 1 - \alpha)$		$\left(\frac{\mathbf{1}(\alpha < e(x) < 1 - \alpha)}{e(x)}, \frac{\mathbf{1}(\alpha < e(x) < 1 - \alpha)}{1 - e(x)}\right)$
overlap	$e(x)(1 - e(x))$	ATO	$(1 - e(x), e(x))$

### 3 Large-sample Properties of Nonparametric Estimators

A natural nonparametric estimator for  $\mathbb{E}(Y|Z = z, X \in dx)$  in a neighborhood  $dx$  of  $x$  is the sample mean of  $Y$  in  $dx$ , denoted by  $\bar{y}_z(dx)$ . Defining  $\hat{\tau}_h(dx) = \bar{y}_1(dx) - \bar{y}_0(dx)$ , we can define a nonparametric estimator for the WATE  $\tau_h$  as follows,

$$\hat{\tau}_h = \int f(x)h(x)\hat{\tau}_h(dx)\mu(dx) / \int f(x)h(x)\mu(dx). \quad (6)$$

Proofs of the following two large-sample results regarding  $\hat{\tau}_h$  are in the Appendix.

**Theorem 1.** *Given the normalizing constraint  $\int f(x)h(x)\mu(dx) = 1$ , as  $N \rightarrow \infty$ , the variance of the estimator  $\hat{\tau}_h$ ,  $\mathbb{V}[\hat{\tau}_h]$ , converges:*

$$N \mathbb{V}[\hat{\tau}_h] \rightarrow \int f(x)h(x)^2 [v_1(x)/e(x) + v_0(x)/(1 - e(x))] \mu(dx), \quad (7)$$

where  $v_z(x)$  is the variance of  $Y$  in a neighborhood  $dx$  of  $x$  in the  $Z = z$  group.

Consequently, if the residual variance is assumed to be homoscedastic across both groups,  $v_1(x) = v_0(x) = v$ , then the asymptotic variance of  $\hat{\tau}_h$  simplifies to

$$N \mathbb{V}[\hat{\tau}_h] \rightarrow v \int \frac{f(x)h(x)^2 \mu(dx)}{e(x)(1 - e(x))}. \quad (8)$$

**Corollary 1.** *The function  $h(x) = e(x)(1 - e(x))$  gives the smallest asymptotic variance for the weighted estimator  $\hat{\tau}_h$  among all  $h$ 's under homoscedasticity, and as  $N \rightarrow \infty$ ,*

$$N \min\{\mathbb{V}[\hat{\tau}_h]\} \rightarrow v \int f(x)e(x)(1 - e(x))\mu(dx).$$

Under mild regularity conditions on  $h$ , the nonparametric estimator  $\hat{\tau}_h$  is equivalent to the standard weighted estimator  $\hat{\tau}_h^w$  with weights normalized in each group (Imbens, 2004):

$$\hat{\tau}_h^w = \frac{\sum_i Y_i Z_i w_1(X_i)}{\sum_i Z_i w_1(X_i)} - \frac{\sum_i Y_i (1 - Z_i) w_0(X_i)}{\sum_i (1 - Z_i) w_0(X_i)}, \quad (9)$$

with the weights  $w$  defined in (4). Therefore, Theorem 1 and Corollary 1 immediately apply to  $\hat{\tau}_h^w$ . In applications, the true propensity score  $e$  is unknown and thus is replaced by the estimated propensity score  $\hat{e}$ . As shown in Rosenbaum (1987) and Hirano *et al.* (2003), a consistent estimate of the propensity score in fact leads to more efficient estimation than the true propensity score.

These theoretical results have some important practical consequences for propensity score weighting analysis. First, calibration of the propensity score model is important: the predicted and empirical rates of treatment assignment should agree in relevant subsets of the covariate space, or else covariate balance cannot be attained. A simple check is to compare predicted and observed rates in deciles (or other convenient intervals appropriate to the amount of data) of the score distribution, in the spirit of Hosmer and Lemeshow (1980). If miscalibration is identified, likely indicating a misfit in the link function, it can be corrected by ad hoc methods such as ratio adjustment of each decile or adding indicator variables for the deciles of a trial model to the final model (if the error is relatively small), or by fitting a flexible logistic spline model relating the estimated linear predictor to treatment assignment. Estimates of treatment effect by decile are sometimes of scientific interest as well for assessing association of treatment *assignment* with treatment *effect*.



Second, a rich propensity score model, rather than a parsimonious one, is desirable, especially in causal inference applications, because the ignorability assumptions are likely to be violated if the propensity score is only a simple logistic-linear function of the covariates but both the outcome and assignment mechanism are functions of more complex interactions or nonlinear terms. Given this, traditional statistical tests of goodness of fit are only minimally relevant, because the objective of weighting is to balance covariate distributions in the *sample*, not to make inferences about assignment probabilities in the *population*. Rather, the limitation on model complexity is imposed by a bias-variance tradeoff. As the model becomes more complex and therefore more predictive, propensity scores move toward 0 and 1, becoming exactly 0 or 1 when the model discovers a separating plane in the data. Thus the weights  $h(x) = e(x)(1 - e(x))$  for many design points move toward zero, reducing the precision of estimates. Analytic judgment is required to decide when further potential reductions in bias are outweighed by the cost in variance. Nonetheless, maintaining the principle of separating design (the weighting model) from analysis (using outcome data), the variance inflation due to model complexity can be approximated by calculating the corresponding ratio for an estimated difference of weighted means assuming homoscedastic data; from (8) this is estimated by

$$(1/n_1 + 1/n_0)^{-1} \sum_{z=0,1} \left( \sum_{i=1}^{n_z} w_{zi}^2 \right) / \left( \sum_{i=1}^{n_z} w_{zi} \right)^2, \quad (10)$$

based on the “design effect” approximation of Kish (1965).

## 4 The Overlap Weights

As Rosenbaum (2012) explains, “... *often the available data do not represent a natural population, and so there is no compelling reason to estimate the effect of the treatment on all people recorded in this source of data...*” In such cases, applying the HT or ATT weights to the

observed sample might not yield an estimate of the average treatment effect of the scientifically appropriate target population. Indeed, the combined or treated population might include individuals that would rarely be considered for one of the treatments, making it meaningless to study the treatment effect including them. For example, a medication might be known to be injurious to patients with certain characteristics. Instead, the natural research interest might lie in the treatment effect on a subgroup who may be exposed to either treatment with substantial probability, that is, whose treatment probabilities are close to neither 0 or 1. In medicine, clinical consensus is ambiguous or divided for such patients, who are said to be in equipoise between treatments, so research on these patients may be most needed. These also might be the units whose treatment assignment would be most responsive to a policy shift as new information is obtained. In these situations, it is natural to “*shift the goalpost*” (Crump *et al.*, 2009) towards estimating effects for the subpopulation in the region of covariate space where distributions for the treatment groups overlap.

We define the *overlap weights* by letting  $h(x) = e(x)(1 - e(x))$ , implying balancing weights

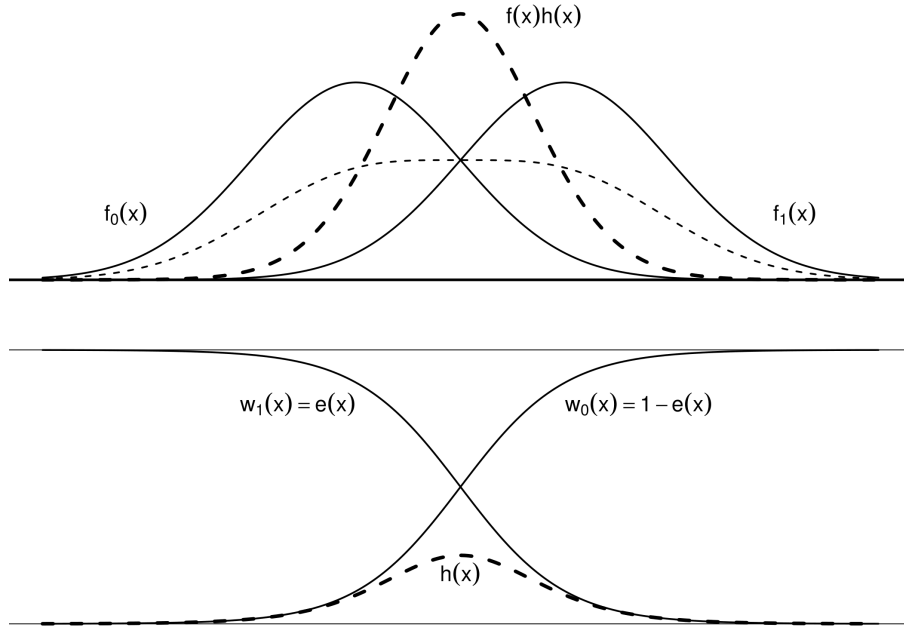
$$\begin{cases} w_1(x) \propto 1 - e(x), & \text{for } Z = 1, \\ w_0(x) \propto e(x), & \text{for } Z = 0. \end{cases} \quad (11)$$

Following Corollary 1 and under its assumptions, the corresponding nonparametric estimator  $\hat{\tau}_h^w$  has the minimum asymptotic variance among all balancing weights.

The overlap weights and the associated target population are illustrated in Figure 1 for two univariate normal populations with equal variance. The upper panel illustrates the target population density  $f(x)h(x)$ , which is greatest where the treated and control groups most overlap. The lower panel shows that the ratio  $h(x)$  of target to combined population peaks where the propensity score is 1/2, and these weights place more emphasis on units with propensity score close to 1/2, who could be in either group. Hence we call them the “overlap weights,” and the corresponding WATE estimand  $\tau_h$  the *average treatment effect for the overlap population*

(ATO). The overlap population thus upweights the potentially policy-sensitive population in equipoise.

Figure 1: Overlap weights for two normally-distributed groups with different means. In the upper panel, the two solid lines and the light and heavy dashed lines represent the density of the covariate in the control, treated, combined ( $h(x) = 1$ ), and overlap weighted ( $h(x) = e(x)(1 - e(x))$ ) populations, respectively. In the lower panel, the two solid lines represent  $w_0(x)$ ,  $w_1(x)$  and the dashed line represents  $h(x)$ .



Statistical advantages of the overlap weights are that they are bounded, unlike the inverse-probability weights, and they are continuously defined and thus avoid arbitrarily truncating weights or excluding units. Moreover, overlap weights based on a logistic propensity score model have an attractive exact (small-sample) balance property. The proof is given in the Appendix.

**Theorem 2.** *When the propensity scores are estimated by maximum likelihood under a*

logistic regression model,  $\text{logit } e(X_i) = \beta_0 + X_i\beta'$ , the overlap weights lead to exact balance in the means of any included covariate between treatment and control groups. That is,

$$\frac{\sum_i X_{ik} Z_i (1 - \hat{e}_i)}{\sum_i Z_i (1 - \hat{e}_i)} = \frac{\sum_i X_{ik} (1 - Z_i) \hat{e}_i}{\sum_i (1 - Z_i) \hat{e}_i}, \quad \forall k, \quad (12)$$

where  $\hat{e}_i = \{1 + \exp(-[\hat{\beta}_0 + X_i\hat{\beta}'])\}^{-1}$  and  $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_k)$  is the MLE for the regression coefficients.

While a main effects model guarantees exact equality between groups for the mean of each included covariate, it is advisable to improve balance by including additional derived covariates, guided by prior anticipation of possible effects on outcomes, as discussed at the end of Section 3. These may include interactions and, for continuous variables, terms whose mean balance implies better matching of distributions, such as powers (to enforce equality of moments) or spline terms.

It is easy to construct examples in which the asymptotic variance (7) of estimates using HT weights is infinite. For example, let  $x \sim \text{Uniform}(0, 1)$ , so  $f(x) = 1$  and let  $e(x) = x$ . Here we illustratively compare HT, truncated HT (retaining cases with  $.1 < e < .9$ ), and overlap weighting estimators under more plausible assumed distributions by calculating variances of WATE estimates relative to the variance of the unweighted difference of means under homoscedastic errors, for selected univariate distributions  $F_1, F_0$  (Table 4). As required by theory, variances with overlap weights are always the smallest. In scenario (1), there is a modest shift of normal distributions; the HT estimator loses efficiency due to extreme weights in the tails, which are excluded by the truncated estimator. With the larger shift between groups in scenario (2), the HT variance is greatly inflated, but the truncated HT weighting removes most of the extreme weights in the tails. In scenario (3) one sample is much larger than the other, which skews the propensity score distribution. This causes an excessive truncation of one tail of that distribution, increasing the variance of the truncated HT estimator. Adaptive modification of truncation points might solve this problem, but no modifications are needed for the overlap

Table 2: Variances of WATE estimators relative to variance of difference of unweighted means, under homoscedasticity and various covariate distributions.

	$F_1$	$F_0$	$n_0/n_1$	Relative variance		
				HT	HT(trunc)	Overlap
(1)	$N(0, 1)$	$N(1, 1)$	1	1.43	1.36	1.26
(2)	$N(0, 1)$	$N(2, 1)$	1	11.81	2.88	2.22
(3)	$N(0, 1)$	$N(1, 1)$	20	2.48	3.31	1.06
(4)	$N(0, 1)$	$N(0, 20^2)$	1	50.02	4.55	3.16

weighting estimator. In scenario (4) one group has much larger variance, again inflating the HT variance, causing an explosion of weights in the tails of the narrower distribution.

The limiting weighting model with a very large number of observations and parameters is a saturated propensity score model with a dummy variable for each design point (or small neighborhood, for continuous variables). In that case, the overlap weights have a natural connection to regression with fixed effects for each design point. If the sample count for  $x_i$  in group  $z = 0, 1$  is  $n_{zi}$ , the propensity score is  $e(x_i) = n_{1i}/(n_{0i} + n_{1i})$  and the total overlap weight for each group and hence of  $\hat{\tau}(x_i)$  is  $n_{0i}n_{1i}/(n_{0i} + n_{1i})$ ; however this is exactly the precision weight attached to  $\bar{y}_{1i} - \bar{y}_{0i}$  in the fixed-effects OLS model  $y_{zi} = \alpha_i + z\tau + \epsilon_{zi}$ .

## 5 Examples

### 5.1 A Descriptive Comparison: Racial Disparities

This application to estimation of racial disparities in medical expenditure uses data similar to that in Lê Cook *et al.* (2010) for adult respondents aged 18 and older to the 2009 Medical Expenditure Panel Survey (MEPS) (AHRQ, 2012). The sample contains 9830 non-Hispanic Whites, 4020 Blacks, 1446 Asians, and 5280 Hispanics. The goal is to estimate disparities in

health care expenditure between non-Hispanic whites and each minority group after balancing confounding variables. We independently conduct three comparisons, comparing each minority group separately to non-Hispanic whites. Since race is not manipulable, these comparisons are descriptive.

There are 27 covariates (5 continuous, 22 binary), a mix of health indicators and demographic variables. We estimate the propensity scores via a logistic regression including all main effects. For each comparison  $Z = 1$  for non-Hispanic Whites and  $Z = 0$  for minority individuals. The distributions of the estimated propensity scores and the normalized overlap and HT weights are shown in Figure 2. The ATT weights are very similar to the HT weights. The White-Asian comparison has the least overlap, and the range of the HT weights in the White-Asian comparison is particularly striking: the largest Asian HT is 0.31, meaning that one individual (out of 1446) contributes over 30% of the Asian estimate in the weighted comparison. This extreme weight belongs to an Asian female with a very high body mass index (BMI) of 55.4 (the highest among Asians). Asians tend to have lower BMI than Whites, so BMI is a strong predictor in the propensity score model, so this individual's propensity score is very close to 1. Thus HT weights give over 30% of the weight for Asians to one atypically obese Asian woman, which is clearly problematic. On the other hand, the largest overlap weight only counts for 0.08%. The common practice to avoid extreme weights when using HT weighting is to truncate these propensity scores, but this can make estimates very dependent on the choice of the cutoff point. For illustrative purposes we continue here with untruncated weights.

Figure 3 provides a closer look at the covariate BMI for the White-Asian comparison, showing the unweighted and weighted distribution of BMI for Whites and Asians under each weighting scheme. This illustrates the good balance achieved by the overlap weights, and the bad balance and extreme weight placed on high BMI for the Asian population under the inverse probability weighting schemes.

Figure 2: Distribution of normalized HT and overlap weights and the estimated propensity scores for whites and each minority group in the MEPS data.

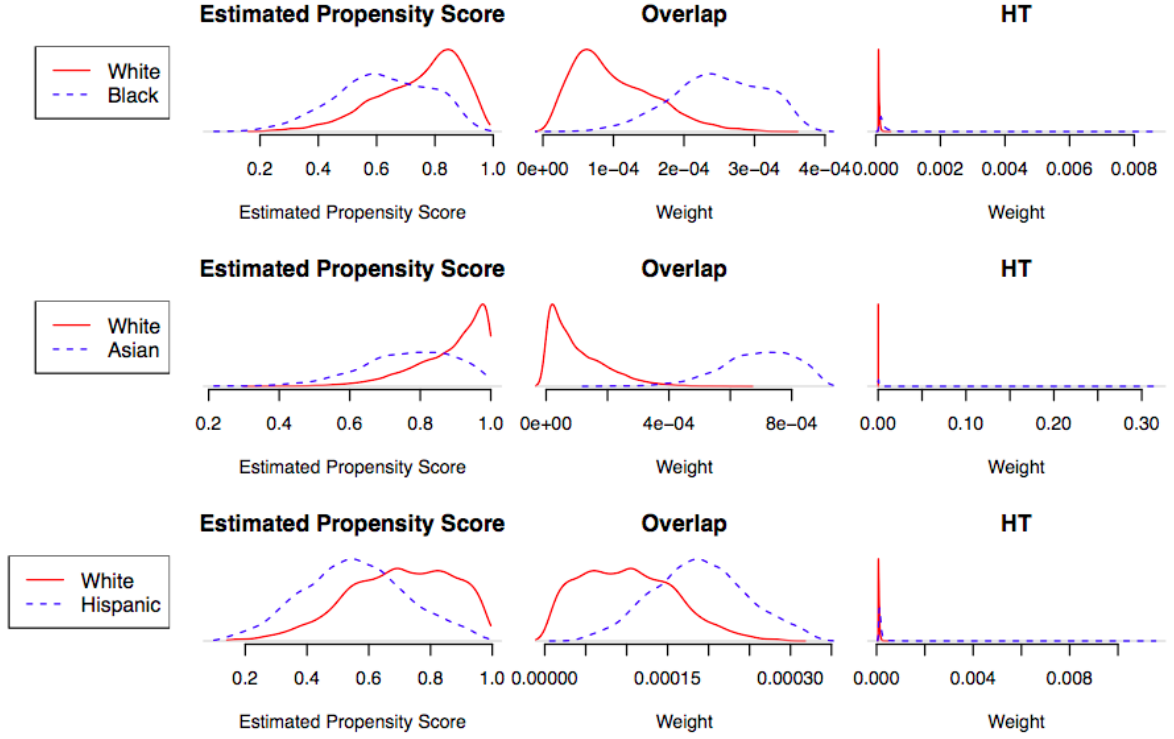
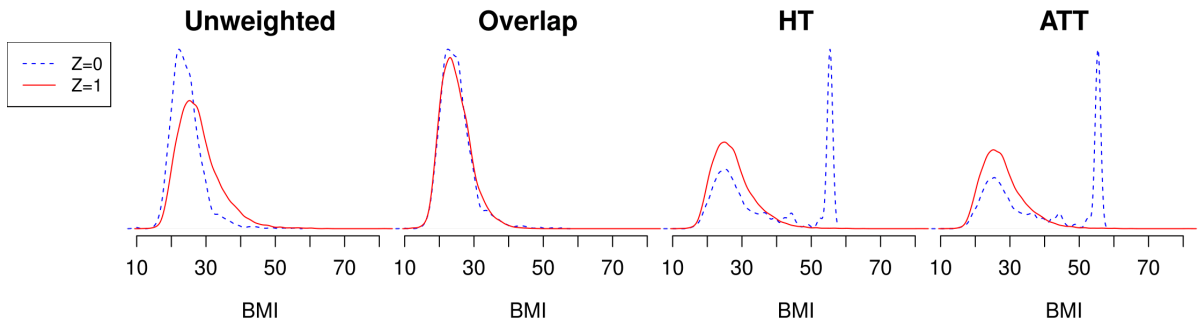


Figure 3: Unweighted and weighted BMI distributions for the White and Asian groups in the MEPS data.



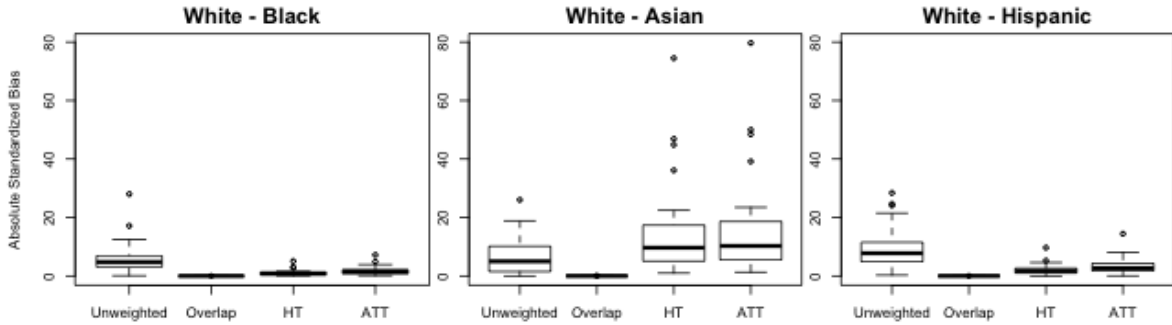
We measure covariate mean balance for each covariate by the absolute standardized bias (ASB), that is, absolute difference in the means of the weighted covariate between Whites and

minority divided by the unweighted standard error:

$$\text{ASB} = \left| \frac{\sum_{i=1}^N X_i Z_i w_i}{\sum_{i=1}^N Z_i w_i} - \frac{\sum_{i=1}^N X_i (1 - Z_i) w_i}{\sum_{i=1}^N (1 - Z_i) w_i} \right| / \sqrt{s_1^2/N_1 + s_0^2/N_0}, \quad (13)$$

where  $s_z^2$  is the variance of the unweighted covariate in group  $z$ . For unweighted data, this is the standard two-sample t-statistic. We use an unweighted standard error in the denominator to allow for fair comparisons across weighting methods. If the numerator and denominator were both to vary with weighting method, a smaller ASB could be due to either better covariate balance or an increased standard error. Figure 4 shows boxplots of the ASB for all covariates under each weighting method. As expected from Theorem 2, the overlap weights lead to perfect balance in all covariates in every comparison. For the White-Black and White-Hispanic comparisons, HT and ATT weighting each substantially improved mean balance compared to the original data, although generally not as well as overlap weighting. For the White-Asian comparison, where there is serious lack of overlap in covariates, HT and ATT weighting performed poorly, yielding very large differences in means for several covariates, including a difference of over 74 standard errors for the covariate BMI. In fact, HT and ATT weighting without truncation results in worse covariate balance than no weighting at all.

Figure 4: Boxplots for the absolute standardized difference for covariates under each weighting method (unweighted, overlap, HT, ATT).



We estimate the weighted average controlled difference in total health care expenditure in



2009 between races using the estimator  $\hat{\tau}_h^w$ . The results with bootstrap standard errors appear in Table 3. Estimates differ substantially across weighting methods, especially when there is a serious lack of overlap, as in the White-Asian comparison. For example, the average difference in total medical expenditure in 2009 between Whites and Asians is estimated to be \$2764, \$1227, \$2167, \$2310 from the un-, overlap-, HT- and ATT-weighted methods, respectively. Overlap weighting focuses on the population where the Whites and minority groups have the most similar characteristics, and gives the smallest standard error in all the comparisons. The HT and ATT estimates for the White-Asian comparison have standard errors more than three times those estimated using overlap weights. In contrast, when there is sufficient overlap, as in the White-Black comparison (first row of Table 3), the estimates and standard errors from different weighting methods are much more similar.

Table 3: Weighted mean differences in total health care expenditure (dollars) in 2009.

	Unweighted (se)	Overlap (se)	HT (se)	ATT (se)
White - Black	786.2 (222.4)	824.3 (184.7)	855.8 (200.3)	851.0 (219.7)
White - Asian	2763.9 (209.5)	1226.7 (204.8)	2167.4 (640.1)	2310.3 (711.1)
White - Hispanic	2598.9 (173.7)	1212.1 (170.7)	596.3 (323.3)	200.2 (445.4)

## 5.2 A Causal Comparison: Right Heart Catheterization

Right heart catheterization (RHC) is a diagnostic procedure for directly measuring cardiac function in critically ill patients. Though useful for directing immediate and subsequent treatment, RHC can cause serious complications. In an influential study Connors *et al.* (1996) used propensity score matching to study the effectiveness of right heart catheterization (RHC) with observational data from Murphy and Cluff (1990). The study collected data on 5735 hospitalized adult patients at five medical centers in the U.S., 2184 of them assigned to the treatment

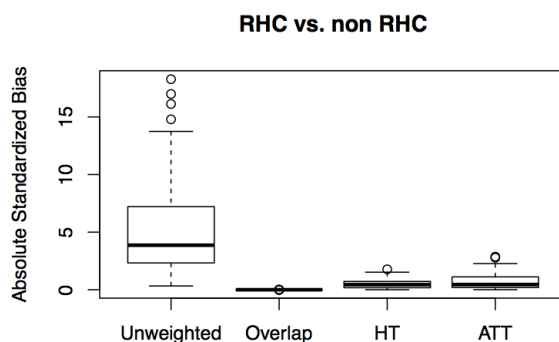
( $Z_i = 1$ ), receipt of RHC within 24 hours of admission, and the remaining 3551 assigned to the control condition ( $Z_i = 0$ ). The outcome was survival at 30 days after admission. Based on information from a panel of experts, a rich set of variables potentially relating to the decision to use RHC was collected. Connors *et al.* (1996) describes the study, which has been since intensively re-analyzed, as in Hirano and Imbens (2001); Crump *et al.* (2009); Traskin and Small (2011); Rosenbaum (2012).

The comparison in the RHC study is causal in the sense that the treatment—application of RHC—is manipulable (Rubin, 1986). Among the 72 observed covariates (21 continuous, 25 binary, 26 dummy variables from breaking up 6 categorical covariates), distributions of several key covariates differed substantially between the control and treatment groups in (Hirano and Imbens, 2001, Table 2)). For example, the treated group has a much higher average APACHE (Acute Physiology and Chronic Health Evaluation) score, signifying greater severity of disease at admission. The majority of the treated units have propensity scores larger than 0.5 and the majority of the control units have propensity scores smaller than 0.5 (Crump *et al.*, 2009, Figure 1). Most previous analyses focus on the ATT, that is, the average causal effect of applying RHC for the patients who received RHC. In this study, arguably it would also be of interest to estimate the effects of RHC for the “marginal” subjects, who might or might not have been treated. Such estimates provide useful information for assigning treatments for the population with no clear propensity to a group in similar later studies. Towards this goal, Rosenbaum (2012) proposed an optimal matching strategy to choose the match with the most treated subjects that has adequate balance. Crump *et al.* (2009) limits the weighting analysis to a subsample with estimated propensity score truncated to the interval  $[0.1, 0.9]$  to ensure good overlap. However, both methods involve dropping some units and the truncation rule depends on the propensity scores, which may be hard to interpret in practice (Traskin and Small, 2011).

As in previous studies, we estimate the propensity score under a logistic model with main

effects of all the 72 covariates, based on which we calculate the HT, ATT and overlap weights. We calculate the ASB for each covariate after applying the overlap, HT, and ATT weights (Figure 5). Each set of weights improves mean balance compared to the unweighted data, and the overlap weights lead to the best balance.

Figure 5: Boxplots for the absolute standardized differences for covariates under each weighting method in the RHC study.



The causal effects estimated from  $\hat{\tau}_h^w$  using the three weights are shown in Table 4. For comparison, we include results from the method of Crump *et al.* (2009) for HT weighting with optimal truncation, and Rosenbaum’s optimal matching method. The standard errors are estimated from the bootstrap procedure in Crump *et al.* (2009) with 5000 replicates. All the estimates suggest that applying RHC leads to a higher mortality rate than not applying RHC. Overlap weighting and optimal truncation lead to the smallest standard errors. Overlap weighting and optimal matching give similar point estimates, around 10% larger than those from the other methods.

Table 4: Estimates and standard errors obtained from different methods. Estimates using optimal truncation are from Crump *et al.* (2009), with estimated propensity score truncated between  $[0.1, 0.9]$  (sample size 4728). Estimates using optimal matching are from Traskin and Small (2011) based on 1563 optimally matched pairs. Standard errors are calculated via bootstrapping.

	unweighted	overlap	HT	ATT	Opt. trunc.	Opt. match
Estimate $\times 10^2$	-7.36	-6.54	-5.93	-5.81	-5.90	-6.78
SE $\times 10^2$	1.27	1.32	2.46	2.67	1.43	1.56

## 6 Discussion and Extensions

Ensuring covariate balance between comparison groups is central to causal and unconfounded descriptive studies. In this paper, we propose a unified framework for the class of weights—the balancing weights—that balance covariates. Several familiar types of weights, such as the HT weight and the ATT weights, are special cases of the balancing weights. Within this general class of weights, we advocate the overlap weights, which optimize the efficiency of comparisons by defining the population with the most overlap in the covariates between treatment groups. This weighting method is easy implement with standard software and has been used in a number of applications, including, for example, Norredam *et al.* (2009) and McWilliams *et al.* (2013) in addition to those mentioned earlier. Though the overlap weights are statistically motivated, we argue that the corresponding target population and estimand are often of scientific or policy interest.

We conclude with a discussion of possible extensions of the weighting methodology.

### 6.1 Multiple Groups

The results of Section 3 readily extend under the same assumptions to optimizing a common target distribution  $f(x)h(x)$  for comparisons of  $J > 2$  groups (Imbens, 2000; Imai and van

Dyk, 2004). Suppose  $e_j(x), j = 1, \dots, J$  are conditional probabilities of assignment to group  $j$ , with  $\sum e_j(x) = 1$ , and that the objective is to minimize the sum of the variances of weighted group means. Then the balancing weights are  $w_j(x) = h(x)/e_j(x)$ , and under the conditions of Corollary 1 they are optimized for  $h(x) \propto (\sum 1/e_j(x))^{-1}$ . Extensions to (known) heteroscedasticity or to an unequally-weighted objective function are straightforward. However, comparison of multiple groups allows consideration of a larger selection of propensity score model specifications than the two-group comparison, as well as different sets of comparisons of interest. (The exact balance results of Section 4 also apply for weights calculated under multinomial or nested logistic models.) Heuristically,  $h(\cdot)$  gives the most relative weight to the covariate regions in which *none* of the  $e_j(\cdot)$  are close to zero. With multiple groups this region of joint overlap might be small or nonexistent, even if all the pairwise overlaps are substantial. Thus, the suitability of weighting to a common distribution depends on the specifics of covariate distributions, models, and scientific objectives of the analysis. Similar issues arise in matching of multiple groups.

## 6.2 Sampling-Weighted Data

When probabilities of selection of cases into the observed sample vary (in particular, within groups), sampling weights are defined as the inverse probabilities of selection. Two basic approaches may be taken to propensity score analysis with sampling-weighted data: ignoring the weights, or incorporating the weights into the propensity-weighting analysis (Zanutto, 2006; DuGoff *et al.*, 2014). In the latter approach, the propensity score model  $e_s(x)$  is estimated using sampling-weighted estimators, and then each observation  $i$  with sampling weight  $W_i$  and corresponding balancing weight  $w_S(x_i)$  is weighted by  $w_S(x_i)W_i$ , where  $w_S(x_i) = h(x_i)/e_s(x_i)$  for  $z_i = 1$ ,  $w_S(x_i) = h(x_i)/(1 - e_s(x_i))$  for  $z_i = 0$ . These approaches have close analogs in regression analysis of sampling-weighted data, in which context use of weights is hotly debated

(Gelman, 2007, and discussion). In either case, the argument for the validity of the weighted estimator is that the weighted sample distribution is an unbiased and consistent estimator of the finite population distribution, and the estimation procedure (regression or propensity score weighting) is consistent when applied to the finite population. On the other hand, a price is paid for use of the sampling-weighted estimator if the combined weights  $w_S(x_i)W_i$  are more variable than the propensity-score weights from the unweighted model  $w_U(x_i)$ , as is likely to be the case, because greater variation in weights typically is associated with increased variance. Furthermore, with weighted data the variance estimator and optimality arguments in Section 3 must be modified because the variance of  $\bar{y}_z(dx)$  is no longer dependent only on sample size, but also on the weight distribution.

The fact that unconfounded comparison depends only on sample balance, not population balance, makes it appealing to ignore the weights and work only with the unweighted samples without explicit reference to the populations from which they are drawn. The problem here is that the unweighted mean  $\bar{y}(dx)$  can only be assumed to be an unbiased estimator of  $\bar{Y}(dx)$  if sampling is ignorable,  $Y \perp W \mid X$ . Otherwise, the weights carry information about the distribution of  $Y$  that is lost in unweighted sample means conditional only on  $X$ , often because the sampling probabilities depend in part on variables that are confidential or too difficult to encode, such as residential address. The plausibility of ignorability conditional on available  $X$  then becomes a crucial issue for validity of unweighted propensity score analysis. A lively stream of current research has sought to devise methods for recovering this information from the weights without the inefficiency of the standard analysis (Zheng and Little, 2005). Application of these approaches to propensity-score weighting could be a productive area for future research.

## 6.3 Weighting and Matching

Since matching is the most widely used nonparametric adjustment method in practice, it is appropriate to compare it to weighting.

Weighting is a “top-down” approach in the sense that it is designed to create global balance for the target population, with increasingly detailed balance appearing as a product of more complex models. Weighting (especially using overlap weights) thus has completely predictable behavior in aggregate.

In contrast, matching is a “bottom-up” approach in the sense that it creates local balance by matching nearby cases, and balance at more aggregated levels is a by-product if the matching is successful. The balance obtained by matching is thus a function of the tuning parameters of the algorithm and features of the covariate distribution, and while balance can be checked, its relationship to these parameters may be somewhat opaque. The compensating benefit is that matching might at least approximately balance distributions in directions orthogonal to the estimated propensity score, improving robustness against deficiencies of the propensity score model. In effect, matching may incorporate high order interactions which could only be incorporated into a propensity weighting model with a very large number of parameters.

Both matching and weighting involve some assumptions of smoothness. Matching analysis assumes that units that are nearby in multivariate space have similar potential outcomes, so the differences between matched treatment and control cases have negligible aggregate effects. In high-dimensional space, however, the distances between nearest matches might still be quite large, and if the distributions of some covariates are very different in the two “treatment” groups, systematic differences might remain after matching. Weighting approaches rely on the more global smoothness of response functions, which makes them close locally to additive models whose approximate sufficient statistics can be balanced by weighting.

Despite these differences, overlap weighting is in a sense asymptotically equivalent to

matching. Consider a sequence of increasingly large datasets from some discrete or continuous generating distribution. The corresponding weighting analyses might be a sequence of increasingly complex models that converge to a model with indicators for each design point or small neighborhood, while the matching criterion of closeness is correspondingly tightened to require exact matching on the same discrete design points or continuous neighborhoods. A many-to-many match at this limit would use weights equivalent to the balancing weights described at the end of section 4. In one-to-one matching, the “excess” observations in a neighborhood from one group might be discarded, yielding estimates from a target distribution  $f(x)h(x) \propto \min(f_1(x), f_0(x))$  which is a form of overlap but uses the data less efficiently than our proposal.

The more parametric specification of weighting gives it more capacity for dealing with some complex data structures. For complex survey data with sample weights, weighting may be more conceptually and practically straightforward than matching, although not without complications, as discussed in the previous subsection. Multilevel data with many small clusters could also pose a challenge to matching. Restricting matches to pairs within the same cluster might not leave enough matching options to obtain good covariate balance, while ignoring the clusters could lead to imbalance on unobserved cluster covariates. With suitable weighting models, cluster and covariate balance can be obtained simultaneously (Li *et al.*, 2013).

The complementarity of these strengths suggests that in some settings a hybrid approach might combine the virtues of matching and weighting: matching followed by an overlap weighting adjustment of the matched sample to eliminate residual imbalance. A similar approach for substitution sampling followed by regression adjustment in surveys was proposed by Rubin and Zanutto (2002).



## Appendix

**Proof of Theorem 1.** Denote the number of units in neighborhood  $dx$  in the  $Z = z$  group by  $n_z(dx)$ , for  $z = 0, 1$ . When the sample size  $N$  increases and the neighborhood  $dx$  decreases,  $n_z(dx)/\{N\mu(dx)\}$  converges to the conditional density of  $x$  in group  $z$ , that is,  $n_1(dx)/\{N\mu(dx)\} \rightarrow f(x)e(x)$  and  $n_0(dx)/\{N\mu(dx)\} \rightarrow f(x)\{1 - e(x)\}$ . Also the variance of the estimated mean outcome for neighborhood  $dx$ ,  $\bar{y}_z(dx)$ , is  $v_z(x)/n_z(dx)$ . Then for the variance of  $\hat{\tau}_h(dx)$ , we have

$$N \mathbb{V}[\hat{\tau}_h(dx)] \rightarrow \{v_1(x)/e(x) + v_0(x)/(1 - e(x))\} / \{f(x)\mu(dx)\}.$$

Given the normalizing constraint  $\int f(x)h(x)\mu(dx) = 1$ , the variance of  $\hat{\tau}_h$  converges,

$$\begin{aligned} N \mathbb{V}[\hat{\tau}_h] &= \int [f(x)h(x)\mu(dx)]^2 N \mathbb{V}[\hat{\tau}_h(dx)] / \left[ \int f(x)h(x)\mu(dx) \right]^2 \\ &\rightarrow \int f(x)h(x)^2 [v_1(x)/e(x) + v_0(x)/(1 - e(x))] \mu(dx). \quad \blacksquare \end{aligned}$$

**Proof of Corollary 1.** Minimizing  $\mathbb{V}[\hat{\tau}_h]$  in (7) subject to the normalizing constraint  $\int f(x)h(x)\mu(dx) = 1$  by Lagrange multipliers gives  $h(x) = e(x)(1 - e(x))$ .  $\blacksquare$

**Proof of Theorem 2.** The score functions of the logistic propensity score model,  $\text{logit}\{e(X)\} = X'\beta$  with  $X_0 \equiv 1$  and  $K$  additional predictors, are:

$$\frac{\partial \log L}{\partial \beta_k} = \sum_i X_i Z_i - \frac{X_i \exp(\beta_0 + X_i \beta')}{1 + \exp(\beta_0 + X_i \beta')}, \quad k = 0, \dots, K.$$

Equating these to 0 and solving, the MLE of  $\beta$  satisfies

$$\sum_i Z_i = \sum_i \hat{e}_i, \quad \text{and} \quad \sum_i X_i Z_i = \sum_i X_i \hat{e}_i,$$

where  $\hat{e}_i = \{X_i \exp(X_i \hat{\beta}')\} / \{1 + \exp(X_i \hat{\beta}')\}$ . It follows that

$$\begin{aligned} \sum_i Z_i (1 - \hat{e}_i) &= \sum_i \hat{e}_i - \sum_i Z_i \hat{e}_i = \sum_i \hat{e}_i (1 - Z_i), \\ \sum_i X_{ik} Z_i (1 - \hat{e}_i) &= \sum_i X_{ik} \hat{e}_i - \sum_i X_{ik} Z_i \hat{e}_i = \sum_i X_{ik} \hat{e}_i (1 - Z_i), \quad k = 1, \dots, K. \end{aligned}$$

Therefore, for any  $k = 1, \dots, K$ , we have

$$\frac{\sum_i X_{ik} Z_i (1 - \hat{e}_i)}{\sum_i Z_i (1 - \hat{e}_i)} = \frac{\sum_i X_{ik} (1 - Z_i) \hat{e}_i}{\sum_i (1 - Z_i) \hat{e}_i}. \quad \blacksquare$$

## References

AHRQ (2012). MEPS HC-132: 2010 Full Year Population Characteristics.

Connors, A., Speroff, T., Dawson, N., Thomas, C., Harrell, F., Wagner, D., Desbiens, N., Goldman, L., Wu, A., Califf, R., Fulkerson, W., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J., and Knaus, W. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *Journal of the American Medical Association*, **276**, 889–897.

Crump, R., Hotz, V., Imbens, G., and Mitnik, O. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, **96**(1), 187–199.

DuGoff, E., Schuler, M., and Stuart, E. (2014). Generalizing observational study results: Applying propensity score methods to complex surveys. *Health Service Research*, **49**(1), 284–303.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, **22**(2), 153–164.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, **66**(2), 315–331.

Hirano, K. and Imbens, G. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, **2**, 259–278.

- Hirano, K., Imbens, G., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, **71**(4), 1161–1189.
- Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.
- Hosmer, D. and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, **A10**, 1043–1069.
- Imai, K. and van Dyk, D. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, **99**(467), 854–866.
- Imbens, G. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, **87**(3), 706–710.
- Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, **86**(1), 4–29.
- Kish, L. (1965). *Survey Sampling*. John Wiley & Sons, Inc., New York.
- Lê Cook, B., McGuire, T., Lock, K., and Zaslavsky, A. (2010). Comparing methods of racial and ethnic disparities measurement across different settings of mental health care. *Health services research*, **45**(3), 825–847.
- Li, F., Landrum, M., and Zaslavsky, A. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, **32**(19), 3373–3387.
- McWilliams, J., Landon, B., and Chernew, M. (2013). Changes in health care spending and quality for Medicare beneficiaries associated with a commercial ACO contract. *J Amer Med Assoc*, **310**(9), 829–836.

- Murphy, D. and Cluff, L. (1990). SUPPORT: Study to understand prognoses and preferences for outcomes and risks of treatments. *Journal of Clinical Epidemiology*, **43**, S1–S123.
- Norredam, M., Meara, E., Landrum, M., Huskamp, H., and Keating, N. (2009). Financial status, employment, and insurance among older cancer survivors. *J Gen Int Med*, **24**(Supp2), S438–S445.
- Robins, J. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, **90**(429), 122–129.
- Rosenbaum, P. (1987). Model-based direct adjustment. *Journal of the Royal Statistical Society: Series B*, **82**, 387–394.
- Rosenbaum, P. (2012). Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics*, **21**(1), 57–71.
- Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**(1), 41–55.
- Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**(1), 688–701.
- Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, **6**(1), 34–58.
- Rubin, D. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, **74**, 318–324.
- Rubin, D. (1980). Comment on ‘Randomization analysis of experimental data: The fisher randomization test’ by D. Basu. *Journal of the American Statistical Association*, **75**, 591–593.

- Rubin, D. (1986). Which ifs have causal answers: comment on ‘Statistics and causal inference’ by P.W. Holland. *Journal of the American Statistical Association*, **81**, 961–962.
- Rubin, D. and Zanutto, E. (2002). Using matched substitutes to adjust for nonignorable nonresponse through multiple imputations.
- Traskin, M. and Small, D. (2011). Defining the study population for an observational study to ensure sufficient overlap: a tree approach. *Statistics in Biosciences*, **3**, 94–118.
- Zanutto, E. (2006). A comparison of propensity score and linear regression of complex survey data. *Journal of Data Science*, **4**, 67–91.
- Zheng, H. and Little, R. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, **21**(1), 1–20.